The Effect of Translationese in Machine Translation Test Sets

Zhang, Mike; Toral, Antonio

Published in:
Proceedings of the Fourth Conference on Machine Translation

Publication date:
2019

The Effect of Translationese in Machine Translation Test Sets

Mike Zhang
Information Science Programme
University of Groningen
The Netherlands
j.j.zhang.1@student.rug.nl

Antonio Toral
Center for Language and Cognition
University of Groningen
The Netherlands
a.toral.ruiz@rug.nl

Abstract

The effect of translationese has been studied in the field of machine translation (MT), mostly with respect to training data. We study in depth the effect of translationese on test data, using the test sets from the last three editions of WMT's news shared task, containing 17 translation directions. We show evidence that (i) the use of translationese in test sets results in inflated human evaluation scores for MT systems; (ii) in some cases system rankings do change and (iii) the impact translationese has on a translation direction is inversely correlated to the translation quality attainable by state-of-the-art MT systems for that direction.

1 Introduction

Translated texts in a human language exhibit unique characteristics that set them apart from texts originally written in that language. It is common then to refer to translated texts with the term *translationese*. The characteristics of translationese can be grouped along the so-called universal features of translation or translation universals (Baker, 1993), namely simplification, normalisation and explicitation. In addition to these three, interference is recognised as a fundamental law of translation (Toury, 2012): "phenomena pertaining to the make-up of the source text tend to be transferred to the target text". In a nutshell, compared to original texts, translations tend to be simpler, more standardised, and more explicit and they retain some characteristics that pertain to the source language.

The effect of translationese has been studied in machine translation (MT), mainly with respect to the training data, during the last decade. Previous work has found that an MT system performs better when trained on parallel data whose source side is original and whose target side is translationese, rather than the opposite (Kurokawa et al., 2009; Lembersky, 2013).

A recent paper has studied the effect of translationese on test sets (Toral et al., 2018), in the context of assessing the claim of human parity made on Chinese-to-English WMT's 2017 test set (Hassan et al., 2018). The source side of this test set, as it is common in WMT (Bojar et al., 2016, 2017, 2018), was half original and half translationese. It was found out that the translationese part was artificially easier to translate, which resulted in inflated scores for MT systems.

Noting that this finding was based on one test set for a single translation direction, we explore this topic in more depth, studying the effect of translationese in all the language pairs of the news shared task of WMT 2016 to 2018. Our research questions (RQs) are the following:

- RQ1. Does the use of translationese in the source side of MT test sets unfairly favour MT systems in general or is this just an artifact of the Chinese-to-English test set from WMT 2017?
- RQ2. If the answer to RQ1 is yes, does this effect of translationese have an impact on WMT's system rankings? In other words, would removing the part of the test set whose source side is translationese result in any change in the rankings?
- RQ3. If the answer to RQ1 is yes, would some language pairs be more affected than others? E.g. based on the level of the relatedness between the two languages involved.

The remainder of the paper will be organized as follows. Section 2 provides an overview of previous work about the effect of translationese in MT. Next, Section 3 describes the data sets used in

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 73–81
Florence, Italy, August 1-2, 2019.

our research. This is followed by Section 4, Section 5 and Section 6, where we conduct the experiments for RQ1, RQ2 and RQ3, respectively. Finally, Section 7 outlines our conclusions and lines of future work.

2 Related Work

There is previous research in the field of MT that has looked at the impact of translationese, mostly on training data, but there are works that have focused also on tuning and testing data sets.

The pioneering work on this topic by Kurokawa et al. (2009) showed that French-to-English statistical MT systems trained on human translations from French to English (original source and translationese target, henceforth referred to as O→T) outperformed systems trained on human translations in the opposite direction (i.e. translationese source and original target, henceforth referred to as T→O). These findings were corroborated by Lembersky (2013), who also adapted phrase tables to translationese, which resulted in further improvements. Lembersky et al. (2012) focused on the monolingual data used to train the language model of a statistical MT system and found that using translated texts led to better translation quality than relying on original texts.

Stymne (2017) investigated the effect of translationese on tuning for statistical MT, using data from the WMT 2008–2013 (Bojar et al., 2013) for three language pairs. The results using O→T and T→O tuning texts were compared; the former led to a better length ratio and a better translation, in terms of automatic evaluation metrics.

Finally, Toral et al. (2018) investigated the effect of translationese on the Chinese→English (ZH→EN) test set from WMT's 2017 news shared task. They hypothesized that the sentences originally written in EN are easier to translate than those originally written in ZH, due to the simplification principle of translationese, namely that translated sentences tend to be simpler than their original counterparts (Laviosa-Braithwaite, 1998). Two additional universal principles of translation, explicitation and normalisation, would also indicate that a ZH text originally written in EN would be easier to translate. In fact, they looked at a human translation and the translation by an MT system (Hassan et al., 2018) and observed that the human translation outperforms the MT system when the input text is written in the original language (ZH), but the difference between the two is not significant when the original language is translationese (ZH input originally written EN). Therefore, they concluded that the use of translationese as the source language in test sets distorts the results in favour of MT systems.

3 Data Sets

We use the test data from WMT16, WMT17, and WMT18 news translation tasks (*newstest2016, newstest2017, and newstest2018*) exclusively, because they provide results using the *direct assessment* (DA) score (Graham et al., 2013, 2014, 2017), which is the metric we will use in our experiments. DA is a crowd-sourced human evaluation metric to determine MT quality. To elaborate, after participants submit their translations produced by their MT systems, a human evaluation campaign is run. This is to assess the translation quality of the systems, and to rank them accordingly. Human evaluation scores are provided via crowdsourcing and/or by participants, using Appraise (Federmann, 2012). Human assessors are asked to rate a given candidate translation by how adequately it expresses the meaning of the corresponding reference translation, thus avoiding the use of the source texts and therefore not requiring bilingual speakers. The rating is done on an analogue scale, which corresponds to an absolute 0-100 scale.

To prevent differences in scoring strategies of distinct human assessors, the human assessment scores for translations are standardized according to each individual human assessor's overall mean and standard deviation score, which is indicated as the z-score in WMT finding papers. Average standardized scores for individual segments belonging to a given system are then computed, before the final overall DA score for that system is computed as the average of its standardized segment scores.

Finally, systems are ranked to produce the shared task results. There is of course the possibility that some systems score similarly in the shared task. If that is the case, those systems are clustered together. Specifically, clusters are determined by grouping systems together, and comparing the scores they obtained. According to the Wilcoxon rank-sum test, if systems do not significantly outperform others, they are in the same cluster, the opposite is the case if they do outperform each other (Bojar et al., 2016, 2017, 2018).

Language Direction	WMT16			WMT17			WMT18		
	# sys.	# seg.	# assess.	# sys.	# seg.	# assess.	# sys.	# seg.	# assess.
Chinese→English				16	32,016	38,736	14	55,734	32,919
English→Chinese				11	22,011	16,253	14	55,734	32,411
Czech→English	12	30,000	16,800	4	12,020	21,992	5	14,915	12,209
English→Czech				14	42,070	32,564	5	14,915	10,080
Estonian→English							14	28,000	28,868
English→Estonian							14	28,000	15,800
Finnish→English	9	63,040	30,080	6	18,012	27,545	9	27,000	18,868
English→Finnish				12	36,024	8,289	12	36,000	9,995
German→English	10	68,800	33,760	11	33,044	36,189	16	47,968	48,469
English→German				16	48,064	10,229	16	47,968	13,754
Latvian→English				9	18,009	30,321			
English→Latvian				17	34,017	6,882			
Romanian→English	7	27,920	16,000						
Russian→English	10	64,960	37,040	9	27,009	24,837	8	24,000	17,711
English→Russian				9	27,009	25,798	9	27,000	27,977
Turkish→English	9	48,640	18,400	10	30,070	25,853	6	18,000	29,784
English→Turkish				8	24,056	2,219	8	24,000	3,644

Table 1: Datasets used in this study (DA scores from WMT16–18 news translation task). Columns contain (from left to right) the number of submitted systems (# sys.), total number of segments prior to quality control (# seg.), and total number of assessments human assessors carried out (# assess.)

Table 1 provides an overview of the number of systems, segments, and assessments in the previously mentioned editions of WMT for all available language directions. These are the datasets that we use in this work.

4 Effect of Translationese on Direct Assessment Scores

The test sets used by Bojar et al. (2016, 2017, 2018) are bilingual, thus having two sides: source text and reference translation. The source is written in the language that is to be translated from (original language), while the reference is written in the language into which the source text is to be translated (target language). In all the test sets used in our experiments English is one of the two languages involved, being either the source or the target.

Taking as an example of WMT test set the one for Chinese-to-English from 2017, this contains 2,001 sentence pairs. Out of these, 1,000 sentences were originally written in Chinese and translated by a human translator into English, hence the target text is translationese. The other half consists of 1,001 sentences that were originally written in English and translated by a human translator into Chinese, hence the source text is translationese in this subset. A graphical depiction of this can be found in Figure 1. The advantage of this procedure is that the same test set can be used for the English-to-Chinese direction, thus reducing the costs involved in creating test sets in half.

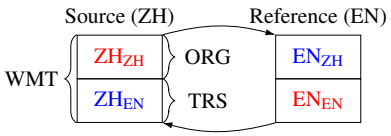

Figure 1: Example of a WMT test set for English (EN) → Chinese (ZH) translation direction, where English is translated into Chinese, and Chinese into English. Indicated as a subscript is which the original language was, red means original language and blue translationese.

Source and reference files contain documents, each of which is provided with a label indicating in which language it was originally written. In our experiments we compute the DA scores for each test set (i) on the whole test set, which corresponds to the results reported in WMT, (ii) on the subset for which the source text was originally written in the source language (referred to as ORG in our experiments) and (iii) on the remaining subset, for which the source text was originally written in the target language, and is thus translationese (referred to as TRS in our experiments).

Table 2 shows the absolute difference in DA score for the ORG and TRS subsets, taking the

Language Direction	WMT16			WMT17			WMT18		
	WMT	ORG	TRS	WMT	ORG	TRS	WMT	ORG	TRS
Chinese→English				73.2	-1.5	+3.9	78.8	-1.3	+2.0
English→Chinese				73.2	-4.1	+5.0	80.7	-4.0	+2.3
Czech→English	75.4	-5.8	+5.7	74.6	-4.3	+4.2	71.8	-1.6	+1.6
English→Czech				62.0	-5.8	+7.4	67.2	-6.6	+7.2
Estonian→English							73.3	-4.0	+4.0
English→Estonian							64.9	-4.1	+3.9
Finnish→English	66.9	-3.2	+3.0	73.8	-2.1	+2.2	75.2	-2.4	+2.3
English→Finnish				59.6	-5.1	+5.6	64.7	-7.7	+8.0
German→English	75.8	-4.1	+4.1	78.2	-2.4	+2.2	79.9	-3.8	+4.3
English→German				72.9	-5.1	+4.4	85.5	-1.9	+1.9
Latvian→English				76.2	-0.4	+0.6			
English→Latvian				54.4	-11.2	+11.7			
Romanian→English	73.9	-0.4	+0.5						
Russian→English	74.2	-1.2	+1.8	82.0	-0.7	+0.6	81.0	-0.1	0.0
English→Russian				75.4	-5.8	+5.8	72.0	-7.4	+7.4
Turkish→English	57.1	-1.6	+1.6	68.8	-3.8	+3.9	74.3	-3.2	+3.9
English→Turkish				53.4	-13.4	+11.8	66.3	-4.1	+5.5

Table 2: DA scores for the best MT system for each translation direction of WMT's 2016–2018 news translation shared task. Columns ORG and TRS show the absolute difference of the DA scores in those subsets compared to the whole test set (WMT).

whole test set (WMT) as starting point for the comparison. We observe a clear and common trend: using original input results in a lower DA score, while using translationese input increases the DA score. This trend is consistent for all the 17 translation directions considered and for all the 3 years of WMT studied, thus providing enough evidence to answer RQ1: the use of translationese as input of test sets results in higher DA scores for MT systems.

5 Effect of Translationese on Rankings

We compute Kendall's τ to give an overview of to what degree rankings change for each translation direction. The τ coefficient is obtained by comparing WMT rankings to the resulting rankings if only the ORG subset is used as input. Since systems can share the same cluster, and thus the same ranking, we compute Kendall's τ both with and without ties. With ties, all systems in the same cluster are considered to occupy the same rank, hence the correlation with ties is sensitive only to changes that go beyond clusters. E.g. if a system moves from the second cluster to the first one. In contrast, without ties all the ranking changes are considered, even if a system changes position but remains within the same cluster.

Table 3 shows the Kendall's τ correlations for all translation directions between the rankings on the whole test set (WMT) and on the ORG subset. We do see that some of the translation directions have a τ coefficient of 1, which means that the agreement between the two rankings is perfect, i.e. the rankings in WMT and ORG are exactly the same. However, we observe that there were few systems submitted to such translation directions (e.g. $\tau = 1$ for Romanian→English in 2017, for which 7 systems were submitted, see Table 1). Apart from those, other language directions show that there are at least slight rank changes between the WMT rankings and ORG rankings. Looking at the low ranked translation directions, we observe that some are close to a τ coefficient of 0, especially in correlations without ties, such as German→English in WMT 2017 ($\tau = 0.345$). This means that some rankings have only a weak correlation.

Probably related to the differences in DA scores between WMT and ORG (RQ1), we also find that systems' rankings change for most language pairs when comparing WMT and ORG rankings. We see that there is no perfect correlation between rankings, apart from a few language directions for which only a few systems were submitted. This

	With Ties			Mean		Without Ties			
Language Direction	WMT16	WMT17	WMT18			WMT16	WMT17	WMT18	Language Direction
Romanian → English†	1.000*	-	-	1.000	1.000	1.000*	-	-	Romanian → English †
Turkish → English	0.983*	0.948*	1.000*	0.977	1.000	1.000*	1.000*	1.000*	Czech → English
Finnish → English	0.943*	0.966*	1.000*	0.970	0.978	-	-	0.978*	English → Estonian †
Czech → English	0.929*	1.000*	0.949*	0.959	0.956	-	-	0.956*	Estonian → English †
German → English	0.979*	0.939*	0.906*	0.941	0.944	-	0.944*	-	Latvian → English †
English → Czech	-	0.904*	0.949*	0.927	0.929	-	0.929*	0.929*	English → Turkish
Latvian → English†	-	0.921*	-	0.921	0.917	-	0.889*	0.944*	English → Russian
English → Finnish	-	0.868*	0.968*	0.918	0.898	-	0.927*	0.868*	English → Chinese
English → Russian	-	0.873*	0.935*	0.904	0.882	-	0.882*	-	English → Latvian †
Chinese → English	-	0.923*	0.882*	0.903	0.869	0.733*	0.944*	0.929*	Russian → English
English → German	-	0.863*	0.856*	0.860	0.852	1.000*	1.000*	0.556*	Finnish → English
English → Estonian†	-	-	0.845*	0.845	0.848	0.833*	0.911*	0.800*	Turkish → English
Estonian → English†	-	-	0.830*	0.830	0.784	-	0.633*	0.934*	Chinese → English
English → Chinese	-	0.847*	0.789*	0.818	0.726	-	0.451*	1.000*	English → Czech
English → Turkish	-	0.890*	0.734*	0.812	0.713	0.911*	0.345	0.883*	German → English
Russian → English	0.557	0.845*	0.890*	0.764	0.675	-	0.817*	0.533*	English → German
English → Latvian †	-	0.718*	-	0.718	0.637	-	0.970*	0.303	English → Finnish

Table 3: Kendall's τ coefficient for each translation direction and year. The coefficient is obtained by comparing WMT's ranking with the ranking if only original language is used as input (subset ORG), with and without ties. A (*) indicates the significance level at p-level p$\leq$0.05. Furthermore, language directions are sorted by the computed mean Kendall's τ. A † indicates that the mean is computed over one year.

Chinese→English

	#	SYSTEM	RAW.WMT	Z.WMT	#	↑↓	SYSTEM	RAW.ORG	Z.ORG	#	↑↓	SYSTEM	RAW.TRS	Z.TRS
wmt17	1	SogouKnowing-nmt	73.2	0.209	1	$2^{\uparrow}$	xmunmt	71.7	0.167	1	$1^{\uparrow}$	uedin-nmt	77.1	0.316
		uedin-nmt	73.8	0.208		$1^{\downarrow}$	SogouKnowing-nmt	71.9	0.161		$1^{\downarrow}$	SogouKnowing-nmt	74.4	0.257
		xmunmt	72.3	0.184		$1^{\downarrow}$	uedin-nmt	70.5	0.101	3	$2^{\uparrow}$	online-A	73.6	0.208
	4	online-B	69.9	0.113		–	online-B	68.7	0.081		$1^{\downarrow}$	xmunmt	72.9	0.202
		online-A	70.4	0.109		$1^{\uparrow}$	NRC	69.1	0.064	5	$1^{\downarrow}$	online-B	71.1	0.145
		NRC	69.8	0.079	6	$1^{\downarrow}$	online-A	67.4	0.012		$1^{\uparrow}$	jhu-nmt	70.0	0.110
	7	jhu-nmt	67.9	0.023	7	–	jhu-nmt	65.8	-0.062		$1^{\downarrow}$	NRC	70.4	0.093
	8	afrl-mitll-opennmt	66.9	-0.016		$1^{\uparrow}$	CASICT-cons	65.4	-0.087		–	afrl-mitll-opennmt	69.2	0.063
		CASICT-cons	67.1	-0.026		$1^{\downarrow}$	afrl-mitll-opennmt	64.5	-0.095		–	CASICT-cons	68.9	0.036
		ROCMT	65.4	-0.058		–	ROCMT	63.4	-0.108		–	ROCMT	67.4	-0.006
	11	Oregon-State-Uni-S	64.3	-0.107		–	Oregon-State-Uni-S	62.7	-0.162		–	Oregon-State-Uni-S	65.9	-0.054
	12	PROMT-SMT	61.7	-0.209	12	$3^{\uparrow}$	online-F	60.0	-0.261	12	–	PROMT-SMT	64.0	-0.137
		NMT-Ave-Multi-Cs	61.2	-0.265		$1^{\downarrow}$	PROMT-SMT	59.4	-0.282		–	NMT-Ave-Multi-Cs	63.3	-0.193
		UU-HNMT	60.0	-0.276		–	UU-HNMT	58.8	-0.301	14	$2^{\uparrow}$	online-G	61.1	-0.245
		online-F	59.6	-0.279		$2^{\downarrow}$	NMT-Ave-Multi-Cs	59.2	-0.337		$1^{\downarrow}$	UU-HNMT	61.1	-0.251
		online-G	59.3	-0.305		–	online-G	57.4	-0.363		$1^{\downarrow}$	online-F	59.2	-0.296
wmt18	1	NiuTrans	78.8	0.140	1	–	NiuTrans	77.5	0.091	1	$8^{\uparrow}$	UMD	80.8	0.239
		online-B	77.7	0.111		–	online-B	77.4	0.089		$6^{\uparrow}$	NICT	80.5	0.232
		UCAM	77.9	0.109		$2^{\uparrow}$	Tencent-ensemble	77.0	0.067		$2^{\downarrow}$	NiuTrans	81.1	0.222
		Unisound-A	78.0	0.108		$1^{\downarrow}$	UCAM	76.3	0.048		–	Unisound-A	80.9	0.222
		Tencent-ensemble	77.5	0.099		$1^{\downarrow}$	Unisound-A	76.4	0.041		$2^{\uparrow}$	Li-Muze	80.7	0.214
		Unisound-B	77.5	0.094		–	Unisound-B	75.8	0.029		$3^{\downarrow}$	UCAM	80.5	0.211
		Li-Muze	77.9	0.091		–	Li-Muze	76.2	0.016		$1^{\downarrow}$	Unisound-B	80.5	0.206
		NICT	77.0	0.089		–	NICT	75.0	0.004		$3^{\uparrow}$	uedin	79.6	0.180
		UMD	76.7	0.078		–	UMD	74.3	-0.021		$4^{\downarrow}$	Tencent-ensemble	78.1	0.149
	10	online-Y	75.0	-0.005		–	online-Y	73.8	-0.047		$8^{\downarrow}$	online-B	78.1	0.147
		uedin	74.5	-0.017	11	–	uedin	71.5	-0.137	11	$1^{\uparrow}$	online-A	77.1	0.068
	12	online-A	73.6	-0.061		–	online-A	71.4	-0.140		$2^{\downarrow}$	online-Y	76.8	0.061
	13	online-G	65.9	-0.327	13	$1^{\uparrow}$	online-F	65.2	-0.353	13	–	online-G	67.8	-0.262
	14	online-F	64.4	-0.377		$1^{\downarrow}$	online-G	64.9	-0.364	14	–	online-F	63.1	-0.417

Table 4: Results of the Chinese→English language direction with WMT, ORG, and TRS input. Systems are ordered by standardized mean DA score. If a system does not contain a rank, this means that it shares the same cluster as the system above it. Clusters are obtained according to Wilcoxon rank-sum test at p-level p $\leq$ 0.05. Indicated in the [↑↓] column are the changes in absolute ranking (i.e. how many positions a system goes up or down).

indicates that the rankings do change to a certain degree. Computing Kendall's τ with ties results in higher correlation coefficients than without ties, implying that systems do shift, but tend to stay in the same cluster they occupied in the WMT ranking. In some editions of WMT, the rankings for certain language pairs change considerably. The biggest change in terms of ranking takes place for PROMT's rule-based system RU→EN for WMT16. This system advances four positions in the ranking when only original source text is considered, going from rank 5 to rank 1 (although tied with several other systems). It is worth noting that while the DA score for the majority of systems decreases when using original source text, the opposite happens for PROMT's system.

Thus far we have looked at a single result per translation direction and year, based on the best system in Table 2, and on the correlation between systems in Table 3. Now we zoom in on a translation direction: Chinese→English. Table 4 shows how DA scores change between the whole test set (WMT) and the subsets ORG and TRS, both in terms of raw and standarized scores. In addition, the table depicts how many positions a system goes up or down in the ranking.

In the table we observe consistently that the DA score for ORG input is lower than that for WMT, while that for TRS is higher than that for WMT. It is also worth noting that most top scoring systems change in rankings, and that system clusters shift. Due to limited space we provide equivalent tables to Table 4 for the remaining 16 translation directions as an appendix.

6 Effect of Translationese on Different Language Pairs

We aim to find out not only whether translationese has an effect on test sets (RQ1 and RQ2), but also to study whether some language pairs are more affected than others (RQ3). Two hypotheses in this regard are as follows: (i) the degree of translationese's impact has to do with the translation quality attainable for a translation direction, as represented by the DA score of the best MT system submitted; (ii) the degree of translationese's impact has to do with how related are the two languages involved.

In order to test the second hypothesis, the degree of similarity between languages has to be quantified. We make use of the lang2vec tool (Littell et al., 2017) using the URIEL Typological Database (Littell et al., 2016) to compute the similarity between pairs of languages. Similar to the approach of Berzak et al. (2017), all the 103 available morphosyntactic features in URIEL are obtained; these are derived from the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013), Syntactic Structures of the Worlds Languages (SSWL) (Collins and Kayne, 2009) and Ethnologue (Lewis et al., 2009). Missing feature values are filled with a prediction from a k-nearest neighbors classifier. We also extract URIEL's 3,718 language family features derived from Glottolog (Hammarström et al., 2019). Each of these features represents membership in a branch of Glottolog's world language tree. Truncating features with the same value for all the languages present in our study, 87 features remain, consisting of 60 syntactic features and 27 family tree features. We then measure the level of relatedness between two languages using the linguistic similarity (LS) by Berzak et al. (2017) (Equation 1), i.e. the cosine similarity between the URIEL feature vectors for two languages v_y and v'_y.

$$LS_{y,y'} = \frac{v_y \cdot v_{y'}}{\|v_y\| \, \|v_{y'}\|} \quad (1)$$

Together with the LS for a language direction, we take the best system of the most recent year in our data set, WMT18, for that language direction. The motivation behind is that a top performing system from the most recent campaign should be representative of the current state-of-the-art in machine translation for the translation direction it was submitted to.

To look into the effect of translationese across different language pairs, we present two approaches, following the hypotheses put forward at the beginning of this section: (i) compare the DA score of the best system for each translation direction on subset ORG to the relative or absolute difference in DA score for that system between subset ORG and the whole set (WMT); (ii) compare the LS of the two languages in each translation direction to the relative or absolute difference in DA scores for the best system between subset ORG and the whole set (WMT);

Figure 2 shows the Pearson correlation and 95% confidence region of the DA score of the best scoring system for each language direction on subset ORG against the absolute and relative difference

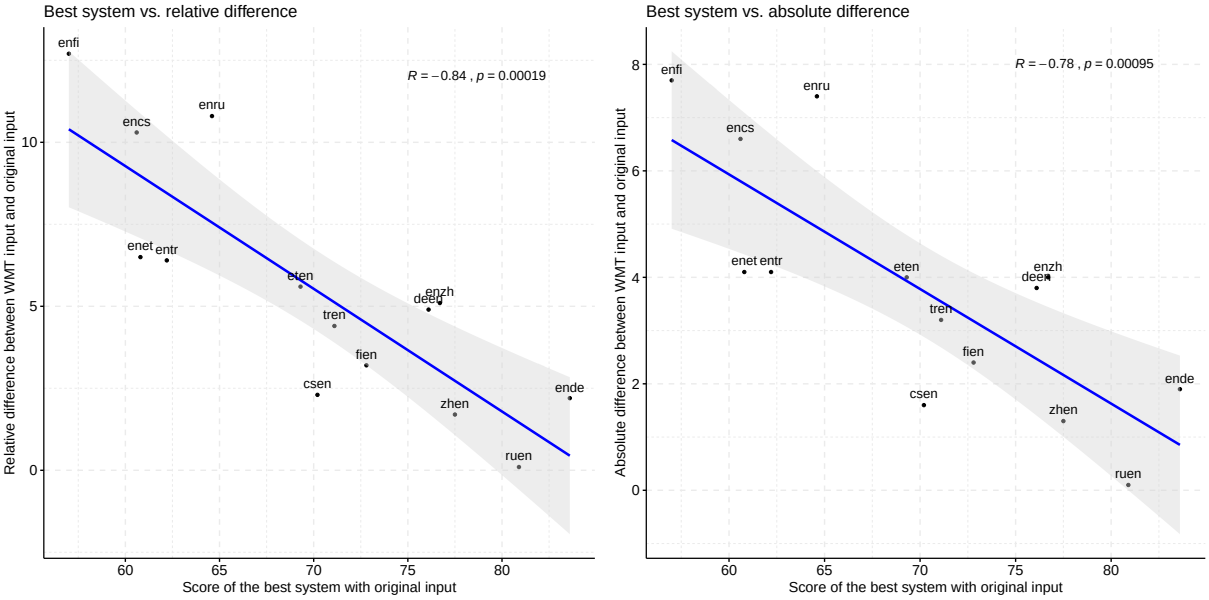

Figure 2: Pearson correlation between the DA scores of the best system for each translation direction at WMT18 and the relative (left) and absolute (right) difference in DA score (%) of comparing WMT input and ORG input. The languages are abbreviated into ISO 639-1 codes (Byrum, 1999).

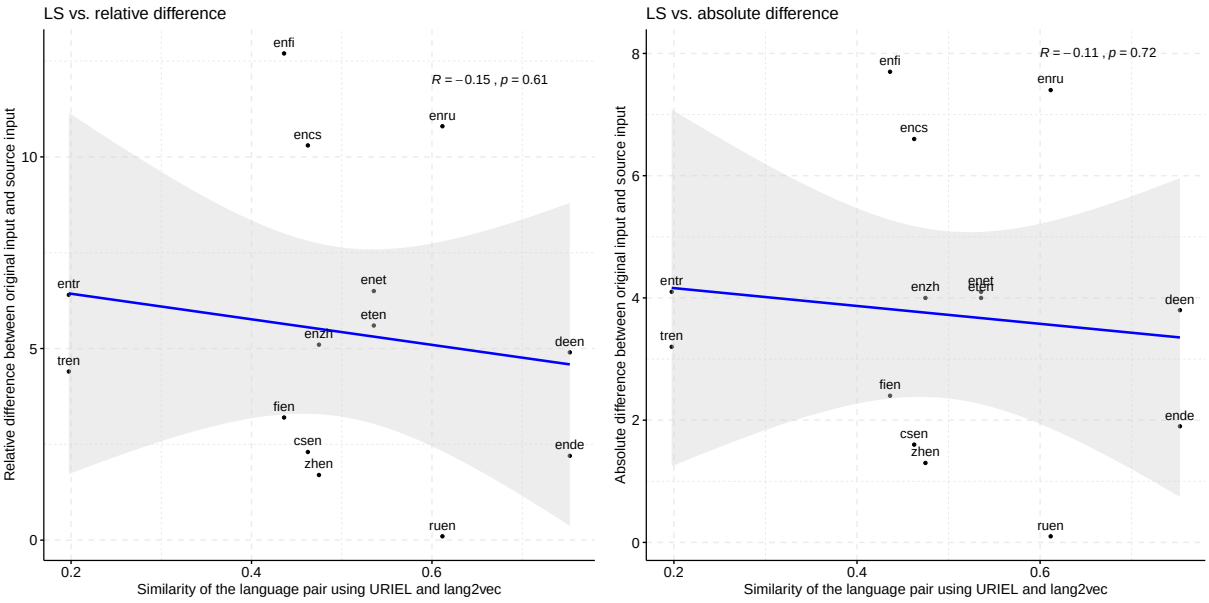

Figure 3: Pearson correlation between Linguistic Similarity for each language direction and the relative (left) and absolute (right) difference (%) in DA score of comparing WMT input and ORG input. The languages are abbreviated into ISO 639-1 codes (Byrum, 1999).

of the DA scores of those systems between WMT input and ORG input. We observe an interesting trend; higher scoring systems tend to have lower differences in score, which indicates that translationese has less effect. Considering either relative or absolute differences, the correlations are in both cases significant and strong ($p < 0.001$, $|R| > 0.75$).

Figure 3 shows the Pearson correlation and 95% confidence region of the LS of a language pair (English compared to another language in our data sets) against the absolute and relative difference of the DA scores of the best system for each translation direction between WMT input and ORG input. Here, we see a less obvious trend, and in fact both correlations are very weak and non-significant. However, just as in the previous figure we can see that most of the out-of-English systems

tend to have a higher relative and absolute difference than systems that translate into English.

On a side note, we created different feature combinations from the earlier mentioned features for LS. Apart from syntactic and family tree features, phonological features are also present in URIEL. However, other combinations did not seem to alter the LS difference score, compared to using the mentioned features in the experimental setup.

7 Conclusion and Future Work

This paper has looked in depth at the effect of translationese in bidirectional test sets, commonly used in machine translation shared tasks, by conducting a series of experiments on data sets for 17 translation directions in the three last editions of the news shared task from WMT. Specifically, we have recomputed the direct assessment (DA) scores separately for the whole test set (WMT), and for the subsets whose source side contains original language (ORG) and translationese (TRS). Results show that using original language input lowers the DA scores, and translationese input increases the scores (RQ1), and perhaps more importantly, system rankings do change (RQ2). We have also investigated the degree to which these rankings change, by measuring the correlation between the rankings with a non-parametric correlation metric that supports ties (Kendall's τ). Results show that systems do change in absolute ranking, but tend to stay more in the same cluster as they were before.

Last, we looked at whether the effect of translationese correlates with certain characteristics of translation directions. We did not find a correlation between the effect of translationese and the level of relatedness of the two languages involved but we did find a correlation between the effect of translationese and the translation quality attainable for translation directions (RQ3). In other words, human evaluation for better performing systems would seem to be less affected by translationese. Related, we observe that translation directions that contain an under-resourced language tend to obtain low DA scores. Hence, we could say that the effect of translationese tends to be high specially when an under-resourced language is present, which could distort (inflate) the expectations in terms of translation quality for these languages.

As for future work, we plan to focus on studying what the characteristics of translationese are. I.e. what are the traits that set apart the language used in original test sets from translationese test sets.

All the code and data used in our experiments are available on GitHub[1].

References

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. *Text and technology: In honour of John Sinclair*, 233:250.

Yevgeni Berzak, Chie Nakamura, Suzanne Flynn, and Boris Katz. 2017. Predicting native language from gaze. *arXiv preprint arXiv:1704.07398*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 131–198.

Ondřej Bojar, Mark Fishel, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Christof Monz, et al. 2018. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation*, pages 272–303.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

John D Byrum. 1999. Iso 639-1 and iso 639-2: International standards for language codes. iso 15924: International standard for names of scripts. In *Proceedings of the 65th International Federation of Library Associations and Institutions (IFLA) Council and General Conference, Bangkok, Thailand.* ERIC.

Chris Collins and Richard Kayne. 2009. Syntactic structures of the worlds languages.

[1] https://github.com/jjzha/translationese

Matthew S. Dryer and Martin Haspelmath. 2013. Wals online. max planck institute for evolutionary anthropology, leipzig.

Christian Federmann. 2012. Appraise: an open-source toolkit for manual evaluation of mt output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2019. Glottolog 3.4. jena: Max planck institute for the science of human history. *Online v.: http://glottolog. org*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

David Kurokawa, Cyril Goutte, Pierre Isabelle, et al. 2009. Automatic detection of translated text and its impact on machine translation. *Proceedings of MT-Summit XII*, pages 81–88.

Sara Laviosa-Braithwaite. 1998. Universals of translation. *Routledge Encyclopedia of Translation Studies. London: Routledge*, pages 288–291.

Gennadi Lembersky. 2013. *The Effect of Translationese on Statistical Machine Translation*. University of Haifa, Faculty of Social Sciences, Department of Computer Science.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. 2009. Ethnologue: languages of the world, dallas: Sil international. *Online version: http://www. ethnologue. com*.

Patrick Littell, David R Mortensen, and Lori Levin. 2016. Uriel typological database. *Pittsburgh: CMU*.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Sara Stymne. 2017. The effect of translationese on tuning for statistical machine translation. In *The 21st Nordic Conference on Computational Linguistics*, pages 241–246.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432*.

Gideon Toury. 2012. *Descriptive translation studies and beyond: Revised edition*, volume 100. John Benjamins Publishing.